DOCUMENT RESUME

ED 374 157 TM 022 086

AUTHOR Kiplinger, Vonda L.; Linn, Robert L.
TITLE Linking Statewide Tests to the National Assessment of Educational Progress: Stability of Results.
SPONS AGENCY National Center for Education Statistics (ED), Washington, DC.
PUB DATE Apr 94
CONTRACT RS90159001
NOTE 46p.; Paper presented at the Annual Meetings of the American Educational Research Association (New Orleans, LA, April 4-8, 1994) and the National Council on Measurement in Education (New Orleans, LA, April 5-7, 1994).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Comparative Analysis; *Educational Assessment; Elementary Secondary Education; *Equated Scores; Estimation (Mathematics); Females; Grade 8; Males; National Surveys; Sex Differences; *State Programs; *Testing Programs; *Test Results
IDENTIFIERS Assumptions (Testing); Coefficients of Stability; Equipercentile Equating; *Linking Metrics; *National Assessment of Educational Progress; Trial State Assessment (NAEP)

ABSTRACT

Recently, several states have expressed interest in linking their statewide assessments to the National Assessment of Educational Progress (NAEP) in the hope that, through equating, they can be compared to national results. This study considers the degree to which existing statewide assessments may be linked to NAEP, without violating the basic assumptions of equating. Results of statewide assessments and of the NAEP Trial State Assessment (TSA) in eighth grade mathematics for both 1990 and 1992 were obtained from four states and equipercentile equating procedures were used. The equating functions for males and females in the two states providing gender identification were similar at the low end of the scale but diverged at the high end of the scale. Estimates of the 1992 NAEP scores derived from applying the 1990 equating functions to the 1992 statewide data were generally similar to actual NAEP results near the median, but were quite dissimilar in the tails of the distribution. These results suggest that such linking, while reasonable for estimating average performance for a state, is not sufficiently stable to use for making comparisons based on the tails of the distribution. (Contains 11 references, 9 tables, and 5 figures.) (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 374 157

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

VONDA L. KIPLINGER

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

LINKING STATEWIDE TESTS TO THE NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS: STABILITY OF RESULTS

Vonda L. Kiplinger and Robert L. Linn
Center for Research on Evaluation, Standards,
and Student Testing
University of Colorado at Boulder

Presented at the Annual Meetings of the American Educational Research Association and the National Council on Measurement in Education, April 4-8, 1994, New Orleans, LA.

BEST COPY AVAILABLE

The research reported herein was supported under the National Center for Education Statistics Contract No. RS90159001 as administered by the U.S. Department of Education. The findings and opinions expressed in this paper do not reflect the position or policies of the National Center for Education Statistics of the U.S. Department of Education.

Abstract

Recently, several states have expressed interest in linking their statewide assessments to the National Assessment of Educational Progress (NAEP) in the hope that, through equating, the results of their own assessments can be compared to national results provided by NAEP. Little is known about the seriousness of violations of conditions assumed for equating. The purpose of this study is to understand better the degree to which existing statewide assessments may be linked to NAEP despite violations of basic underlying assumptions of equating. Results of statewide assessments and of the NAEP Trial State Assessments (TSA) in eighth grade mathematics for both 1990 and 1992 were obtained from four states and equipercentile equating procedures were used. The equating functions for males and females in the two states providing gender identification were similar at the low end of the scale but diverged at the high end of the scale. Estimates of 1992 NAEP scores derived from applying the 1990 equating functions to the 1992 statewide data were generally similar to actual NAEP results near the median, but were quite dissimilar in the tails of the distribution. These results suggest that such linking, while reasonable for estimating average performance for a state, is not sufficiently stable to use for making comparisons based on the tails of the distribution.

Introduction

During the past few years there has been considerable discussion among educational policymakers and measurement specialists regarding the possibility of linking data from different assessments. In addition, several states have expressed an interest in linking their statewide assessments to the National Assessment of Educational Progress (NAEP). There also is a desire to link NAEP to international assessments such as the 1991 International Assessment of Educational Progress (IAEP) (Lapointe, Mead, & Askew, 1992) or the Third International Mathematics and Science Study (TIMSS) that is planned for 1995 (International Association for the Evaluation of Educational Achievement, 1992). It is hoped that through linking, the results of a state's own assessment can be compared to national results provided by NAEP and possibly even to international results through a linking of NAEP to IAEP or TIMSS.

It has long been common practice to equate results of different forms of a test and then treat the results from administrations of different forms as interchangeable. For example, different forms of college admissions tests are given on different administration dates for reasons of test security, but because the scores on the different forms have been equated the results can be treated as if a single form of the test had been administered. In a similar fashion, achievement test publishers routinely publish alternate forms of an achievement test that are equated to a common scale so that users can obtain comparable results using a particular form one year and another form the next year.

As has been discussed by a number of authors, the claim that two test forms have been equated is a strong one and stringent criteria must be satisfied if the claim is to be defensible (cf. Linn, 1993; Lord, 1980; Mislevy, 1992). The claim implies that the test form administered should be a matter of indifference to anyone taking the test and to anyone using the results. This indifference property of equated test forms is important for the equitable use of the results. Though never perfectly realized in practice, it can be reasonably approximated, but only if certain conditions are satisfied. As Porter (1991: 35) has stated quite clearly, "Equating can be done only when tests

measure the same thing". In addition, the tests must measure the domain in question with equal precision.

Even tests that are designed with such constraints in mind only approximate the stringent conditions. Tests or assessments constructed for different purposes using different content frameworks or specifications will almost surely violate the conditions required for a strict equating. A question remains, however, whether sufficiently trustworthy results can be obtained by using either statistical equating procedures or some other statistical procedure designed to serve more modest goals.

Types of linking that have less stringent requirements and, in turn, yield weaker results that support comparisons in more limited circumstances are discussed by Linn (1993) and by Mislevy (1992). We will not review those distinctions here, but simply note that validity of comparisons across tests or assessments may depend on the context of assessments, the groups used to calculate statistics, and the time of administration. For example, an equation that would enable a state to use its statewide assessment to predict with reasonable accuracy the results that would be obtained on NAEP in one year might yield quite inaccurate results in another year.

Although the theoretical restrictions on equating are well known, there is less empirical information regarding the seriousness of violations of conditions assumed for equating of the type that may be encountered with the actual assessments that educational policymakers would like to have linked. The purpose of this study is to add to the available empirical results to provide a better understanding of the degree to which existing statewide assessments may be linked to NAEP despite violations of basic underlying assumptions that the assessments are measuring the same construct with equal precision.

Related Studies

Two recent studies have attempted to link either the 1990 or 1992 NAEP mathematics assessment to the 1991 IAEP mathematics assessment (Beaton & Gonzalez, 1993; Pashley & Phillips, 1993). The results obtained in these two studies were similar for countries with average performance near that of U.S. students. For

example, eighth grade students in Spain had an average percent correct score equal to that of U.S. students (55) on the IAEP mathematics assessment. The estimates of the percentage of students in Spain expected to exceed 294 on the NAEP scale (the minimum score for the "proficient" achievement level) were 10.7 percent in the Beaton and Gonzalez analysis and between 10.4 and 13.0 percent in the Pashley and Phillips analysis.

For countries with very high performance on the IAEP, for example, Taiwan and Korea (both with average percent correct scores on the IAEP of 73, as opposed to 55 for U.S. students), the two analyses yielded quite discrepant results. The estimate of the percentage of students who would score above 294 on the NAEP scale in Taiwan was 54.1 in the Beaton and Gonzalez analysis, compared to between 34.6 and 39.3 in the Pashley and Phillips analysis. The corresponding figures for Korea were 52.2 percent, compared to between 38.2 and 43.1 percent. Using a higher cut score of 331 (the minimum score for the "advanced" achievement level) results in a even larger discrepancy. Beaton and Gonzalez estimated that 24.4 percent of students in Taiwan performed at this advanced level, whereas Pashley and Phillips estimated that only between 5.3 and 7.5 percent were at that level. The results obviously are sensitive to differences in the data bases and the techniques used to link IAEP results to NAEP.

Another recent study (Ercikan, 1993) is more closely related to the present one. Ercikan used equipercentile equating procedures (cf. Peterson, Kolen, & Hoover, 1989) to convert statewide results on one of the standardized tests published by CTB Macmillan/McGraw-Hill into predicted performance on the 1990 NAEP scale. Data were obtained from four states that participated in the NAEP Trial State Assessment (TSA) in mathematics at grade 8 in 1990. Data also were obtained from statewide administrations of the California Achievement Tests Form E (CAT/E), administered in one state, and the Comprehensive Tests of Basic Skills Form E or Form 4 (CTBS/E or CTBS/4), administered in four states.

The CAT and CTBS scores were first converted to the Normal Curve Equivalent (NCE) scale of the CAT/5, which is the latest edition of the CAT. The resulting NCE scores for the standardized tests were then converted to the NAEP scale using an

equipercentile equating procedure. Within-state equatings were performed using the results from each individual state. In addition, an equating was performed for the combined data from the three states using one of the CTBS forms. Finally, an equating was performed using the combined data from all four states.

If the conditions for equating were fully satisfied, the results of the six equatings would be expected to be identical except for sampling error. However, the results showed considerably greater divergence than would be expected due to sampling error alone. For example, a NCE score of 90 on the CAT predicted NAEP scores ranging from a low of 305 in one state to a high of 325 in another state. Twenty points on the NAEP mathematics scale corresponds to almost two-thirds of a standard deviation for the national sample at grade 8. Although not presented by Ercikan, standard errors of equating for samples of the size used would be roughly only one to two points.

One likely reason for the divergence of results among the different states is that NAEP and the standardized tests do not measure the same thing. A recent investigation of the content convergence between NAEP and three standardized mathematics tests at grade 8 was conducted by Bond and Jaeger (1993) to evaluate that possibility. One of those tests, the CAT, was used by both Ercikan and by one of the states of the present study. A second test analyzed by Bond and Jaeger, the Stanford Achievement Test (SAT), also was used by two of the states participating in the present study.

Bond and Jaeger enlisted the assistance of a group of content experts in mathematics to independently classify items from each of the standardized tests into one of the NAEP subject-matter categories or into an "unclassifiable" category. The five subject-matter categories are Numbers and Operations; Measurement; Geometry; Data Analysis, Statistics, & Probability; and Algebra & Functions. The judges also classified the standardized test items according to the three "ability" categories of the NAEP framework (Conceptual Understanding, Procedural Knowledge, and Problem Solving). The results indicated that a disproportionately large number of items from all three standardized tests were classified into either the Numbers and Opera-

tions/Procedural Knowledge category or the Numbers and Operations/Conceptual Understanding category. The Bond and Jaeger results for the CAT and SAT are quite relevant to the present study and will be discussed in greater detail below.

Methodology

The present study is similar to the Ercikan study in that statewide results for standardized tests, together with NAEP-TSA results, were obtained from four states and equipercentile equating procedures were used. The present study differs in the standardized tests used and, more importantly, in that data were obtained for both 1990 and 1992. Having data from two statewide assessments in grade 8 mathematics and two administrations of the NAEP-TSA makes it possible to obtain an equating function that converts the statewide results in 1990 to the 1990 NAEP-TSA results and then to evaluate the accuracy of the conversion when the equating function is applied to data collected in 1992.

Data Sources

Data from statewide administrations of standardized tests in grade 8 mathematics in 1990 and 1992 were obtained from four states that participated in both the 1990 and 1992 Trial State Assessments in mathematics at grade 8. The standardized tests used each year and the sample size for the four states providing data for this study are listed in Table 1. As can be seen, two states used different forms of the Stanford Achievement Test (SAT), one state used the Iowa Tests of Basic Skills, and one used the California Achievement Test.

Table 1

The number of years that a particular standardized test form had been used varied among the four states. Form K of the SAT was used for the first time in 1990 and the third time in 1992 in State 1. Form L of the SAT was administered for the

first time in 1992 in State 2. Prior to that time, Form E of the SAT had been used for several years. In both States 3 and 4, the 1990 data collection was the fifth year of administration of their test forms, while the 1992 data collection was the seventh. These varied patterns are potentially relevant since test scores tend to show a decline the year a new test form is introduced and then increase most rapidly during the next two or three years of use, with small or negligible changes in subsequent years (cf. Linn, Graue, & Sanders, 1990).

Analyses

The 1990 statewide test results and the 1990 TSA results were used in the main equating analyses. For the NAEP-TSA, the average percentile values were obtained from the NAEP contractor, Educational Testing Service. Those percentiles are based on estimations from the five plausible values used in NAEP statistical analyses and take the sampling weights and complex sample design into account to produce estimates for a state. The percentiles for the statewide assessments were computed using the scaled scores that were provided by the states. Since the statewide test administrations are intended to be a census, the use of sampling weights was not required to obtain statewide results.

The standardized test results were converted to the NAEP scale using the 1990 data. As is illustrated in Figure 1, the resulting conversion tables were then applied to the 1992 results on the statewide test to obtain estimated 1992 results for the state on NAEP. The estimated NAEP results were then compared to the actual NAEP scores obtained in the 1992 TSA administration. For State 2, where different forms of the SAT were used in the two years, the 1992 SAT results were first expressed in terms of the 1990 SAT scale using conversion tables provided by the state; those results then were converted to the NAEP scale in same manner as the other states.

Figure 1

Results

If all conditions required for equating are satisfied, then, except for sampling error, the equating functions should be invariant across subpopulations (e.g., males and females). However, the results obtained in this study for the two states providing gender identification yield differences larger than would be expected based on sampaling error alone for some parts of the distrubutions.

The equating functions for the 1990 SAT Total Mathematics and the NAEP Overall Proficiency scores using the data from State 1 are displayed in Figure 2 for the state total and for males and females. As can be seen in Figure 2, a given score on the SAT would be converted to a somewhat higher score on the NAEP if the equating function for males was used rather than the equating function for females. Also, the difference between the two equating functions tends to be larger at the low end of the distribution than at the high end.

Figure 2

The magnitude of the difference at selected percentile points for the total group from State 1 is shown in Table 2. Columns two and three of Table 2 list the SAT Total Mathematics and the NAEP Overall Mathematics Proficiency scores corresponding to total group percentiles of 95, 90, 75, 50, 25, 10, and 5. Estimated NAEP scores based on the separate male and female equating functions are shown in columns four and five. Finally, the difference between the putatively equivalent scores from the male and female equatings are shown in column six.

Table 2

If all conditions that are required for equating are satisfied, then, expect for sampling error, the equating functions should be invariant across subpopulations. Approximate standard errors of equating were computed for various percentiles using the formula given by Petersen, Kolen and Hoover (1989:251) for the two-group equipercentile case. For State 1, the standard error of equating for males or females varies from a low of approximately 1.1 points at the 50th percentile to a high of approximately 1.9 points at the 5th and 95th percentiles. The standard error of the difference for the independent samples ranges from about 1.6 at the 50th percentile to approximately 2.6 at the 5th and 95th percentiles. Thus, as shown in Table 2, the differences for all but the 95th percentile are more than twice their standard errors.

The equating functions for the NAEP Overall Proficiency scores and total SAT math scores for State 2 are presented in Figure 3. As shown in Figure 3, the equating functions for State 2 are similar to those for State 1 in that a given score on the SAT generally would be transformed to a slightly higher score on the NAEP if the equating function for males rather than the function for females was used. Also, the differences between the male and female equating functions are larger at the lower end of the distributions. However, unlike the functions for State 1, the differential all but disappears for SAT scores of 91 or higher.

Figure 3

The differences between the male and female equatings of the SAT and NAEP average overall proficiency scores at selected percentiles for State 2 are presented in Table 3. As is indicated, only the differences at the 10th and 25th percentiles exceed twice their standard errors. Also shown in Table 3 are the SAT scores and the equivalent NAEP scores corresponding to the selected percentiles (95, 90, 75, 50, 25, 10, and 5) for the total group.

Table 3

The content analyses conducted by Bond and Jaeger (1993) suggested that the majority of the items on the SAT belong to one of the five NAEP content areas -- Numbers and Operations. Consequently, separate equipercentile equatings were performed using the NAEP Numbers and Operations scores, rather than the Overall Mathematics Proficiency scores, and the SAT Total Mathematics scores as before. The results of those equatings are shown in Figure 4 and Table 4 for State 1 and in Figure 5 and Table 5 for State 2.

The equating functions relating the SAT Total Mathematics scores and NAEP Numbers and Operations scores for State 1, presented in Figure 4, are very similar to those relating the SAT Total Mathematics scores and NAEP Overall Mathematics Proficiency scores shown in Figure 2. As illustrated in the figures and shown by comparison of Tables 2 and 4, the equating functions for males and females are most divergent at the low end of the distribution when either the NAEP Numbers and Operations scores or the NAEP Overall Mathematics Proficiency scores were used. The differences are greater than twice their standard errors for scores corresponding to the 75th percentile or lower.

Figure 4 and Table 4

As with State 1, the equating functions relating the SAT Total Mathematics scores and NAEP Numbers and Operations scores for State 2 are very similar to those relating the SAT Total Mathematics scores and NAEP Overall Mathematics Proficiency scores. These equating functions are presented in Figures 5 and 3, respectively. Comparison of Tables 3 and 5 also indicates that the equating functions for males and

females in State 2 are more divergent at the lowest reported percentile (5th) in the equating using the NAEP Numbers and Operations scores than in the equating with the NAEP Overall Mathematics Proficiency scores. Otherwise, the results for the male-female differences are reasonably similar for the two different NAEP scores.

Figure 5 and Table 5

Gender identification was not available for the statewide test data provided by States 3 and 4. Hence, there is no check on the total group equating from the 1990 data alone. For all four states, however, the primary check on equating is based on the application of equating functions derived from the 1990 data to the data obtained in 1992.

The scores on the statewide tests corresponding to percentiles of 5, 10, 25, 50, 75, 90, and 95 in 1992 were obtained in each state. Those statewide test scores were then converted to estimates of the corresponding 1992 NAEP scores using the 1990 equating functions. The resulting estimates of the 1992 NAEP scores were then compared to the 1992 NAEP scores that were actually observed the Trial State Assessment for those selected percentiles for each state.

Table 6 lists the results comparing estimated and observed 1992 NAEP Overall Proficiency scores for State 1. In general, the differences between estimated and obtained scores were reasonably small. Only at the low end of the distribution (5th and 10th percentiles) did the differences exceed two standard errors.

Figure 6

The results of the comparison of estimated and observed NAEP Overall

Proficiency scores for State 2 are shown in Table 7. Since a new form of the SAT was used in 1992, scores on the new form first had to be equated to the scale of the form used in 1990 and then mapped into the NAEP scale using the 1990 SAT to NAEP conversion. As can be seen in Table 7, estimated and observed performance on NAEP was similar for the bottom half of the distribution; however, the observed performance was higher than the estimated performance for the top half of the distribution.

Table 7

A comparison of the estimated and observed 1992 NAEP Overall Proficiency scores for State 3 is presented in Table 8. In this state, the observed NAEP scores are higher than those estimated by the equating function, particularly at or above the 75th percentile. That is, equipercentile equating underestimates the 1992 NAEP Overall Proficiency scores in mathematics in State 3.

Table 8

The estimated and observed 1992 NAEP Proficiency scores for State 4 are compared in Table 9. This table indicates that the 1992 NAEP Overall Mathematics Proficiency scores are substantially over-estimated by the equipercentile equating procedure, particularly above the median and at the 5th percentile.

Table 9

Discussion

If the conditions required for equating are completely satisfied, then equating functions for different subgroups (e.g., males and females) should be the same except for sampling error. The results obtained in this study for the two states where gender identification is available yield differences larger than would be expected based on sampling error alone for some parts of the distributions. The differences in the region between the 5th and 95th percentiles are as large as 11 points for State 1 and 8 points for State 2.

Results from the content analyses reported by Bond and Jaeger (1993) suggest that the failure to obtain essentially the same equating functions for different subgroups may be due to differences in the content coverage of the NAEP and the statewide tests. Given their analysis, one might expect that the equating functions would be more similar when the statewide tests are equated to the Numbers and Operations scale than when equate to the Overall Mathematics Proficiency scale. The differences in the male and female equating functions are of similar magnitude for the two types of NAEP scales, however.

The main comparisons of this study focused on the accuracy of the estimates when 1990 equating functions were used with 1992 statewide test data to estimate the 1992 NAEP results. These comparisons reveal differences that are larger than expected, based on sampling error alone, in one or both tails of the distribution in all four states. If conditions required for equating are completely satisfied, then any changes in the mathematics achievement of students within a state between 1990 and 1992 should have comparable effects on both NAEP and the statewide test results and, therefore, the equating obtained with 1990 data should still hold in 1992.

The obtained differences between the estimated and actual 1992 results indicate that there are violations of assumptions required for a strict equating in all four states. For some restricted purposes, however, the differences might be considered to be acceptably small. Results at or near the median, for example, were small for three of the four states. Consequently, the linking might be considered adequate for purposes of estimating average achievement on the NAEP scale, but not

for estimating achievement at the lower or upper ends of the distribution.

For two of the states, the magnitude and sign of the differences between actual and estimated 1992 performance on NAEP varied in accord with what might be expected from the length of time a particular form had been used in each state. State 1, where observed scores were lower than estimated, administered the standardized test form for the first time in 1990 and the third time in 1992. Previous research (e.g., Linn, Graue, & Sanders, 1990) has shown that relatively large increases are frequently observed between the first and second or third year of test administration. To the extent that gains during the first few years that a new form is used are the result of increased familiarity with and emphasis on the specific content of the test, one would expect that the gains would not generalize to other measures such as NAEP. This expectation is consistent with the results of over-estimation of NAEP scores obtained for State 1.

In State 2, where a new form was used for the first time in 1992, results show the opposite pattern. That is, for the upper half of the distribution, the observed NAEP scores are higher than the estimated scores. The commonly observed decline in scores when a new form is first introduced provides a plausible explanation of this finding. That is, the apparent dip in performance on the standardized test is largely an artifact of somewhat inflated results in 1990 due to the repeated use of the old form. Neither NAEP nor the new standardized test form is subject to that inflation. Hence, the equating function derived in 1990 leads to underestimates of NAEP performance in 1992 when it is applied to the 1992 standardized test results.

Both States 3 and 4 used a standardized test form for the fifth time when they were administered in 1990 and for the seventh time in 1992. Whatever inflation in test scores that is due to familiarity with and emphasis on test-specific content is likely to have been realized by the fifth administration. Thus, there seems to be little reason to expect the estimates or 1992 NAEP scores to be either too high or too low and we lack any substantive hypothesis as to why the NAEP scores tended to be underestimated in State 3 and overestimated in State 4, especially at the higher end of the distributions.

No matter what the substantive explanation for the lack of stability of the equating functions from 1990 to 1992, it seems clear that there is substantial uncertainty in the estimates. The lack of stability suggests that linking standardized tests to NAEP using equipercentile equating procedures is not sufficiently trustworthy to use for other than rough approximations. In considering the results of this study, however, it should be recalled that these tests were not designed with the purpose of linking in mind. The content differences between the standardized tests and the NAEP framework identified by Bond and Jaeger (1993) are substantial. More stable results might be expected if the tests being linked were designed in accordance with a common framework. If linking is an important goal, then it would seem wise to assure, at a minimum, that the tests share a common content framework.

References

Beaton, A. E. & Gonzales, E. J. (1993). Comparing the NAEP Trial State Assessment results with the IAEP International results. Report prepared for the National Academy of Education Panel on the NAEP Trial State Assessment. Stanford, CA: National Academy of Education.

Bond, L. & Jaeger, R. M. (1993). Final report on the judged congruence between various statewide assessment tests in mathematics and the 1990 National Assessment of Educational Progress. Report prepared for the National Academy of Education Panel on the NAEP Trial State Assessment. Stanford, CA: National Academy of Education.

Ercikan, K. (1993). Predicting NAEP. Unpublished manuscript. Monterey, CA: CTB Macmillan/McGraw-Hill.

International Association for the Evaluation of Educational Achievement. (1992). The Third International Mathematics and Science Study: Project Overview. Vancouver, BC:IEA TIMSS International Coordinating Center.

Lapointe, A. E., Mead, N. A., & Askew, J. M. (1992). Learning mathematics. Princeton, NJ: Educational Testing Service.

Linn, R. L., Graue, M. E., & Sanders, N. M. (1990). Comparing state and district test results to national norms: The validity of claims that "everyone is above average". Educational Measurement: Issues and Practice, 9, 5-14.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Mislevy, R. J. (1992). Linking educational assessments: Concepts, issues, methods, and prospects. Princeton, NJ: Educational Testing Service, Policy Information Center.

Pashley, P. J. & Phillips. G. W. (1993). Toward world-class standards: A research study linking international and national assessments. Princeton, NJ: Educational Testing Service.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), Educational measurement (3rd Ed., pp. 221-262). New York: Macmillan.

Porter, A. C. (1991). Assessing national goals: Some measurement dilemmas. In T. Wardel (Ed.), The assessment of national goals: Proceedings of the 1990 ETS Invitational Conference. Princeton, NJ: Educational Testing Service.

Figure 1

Schematic Representation of Study Design

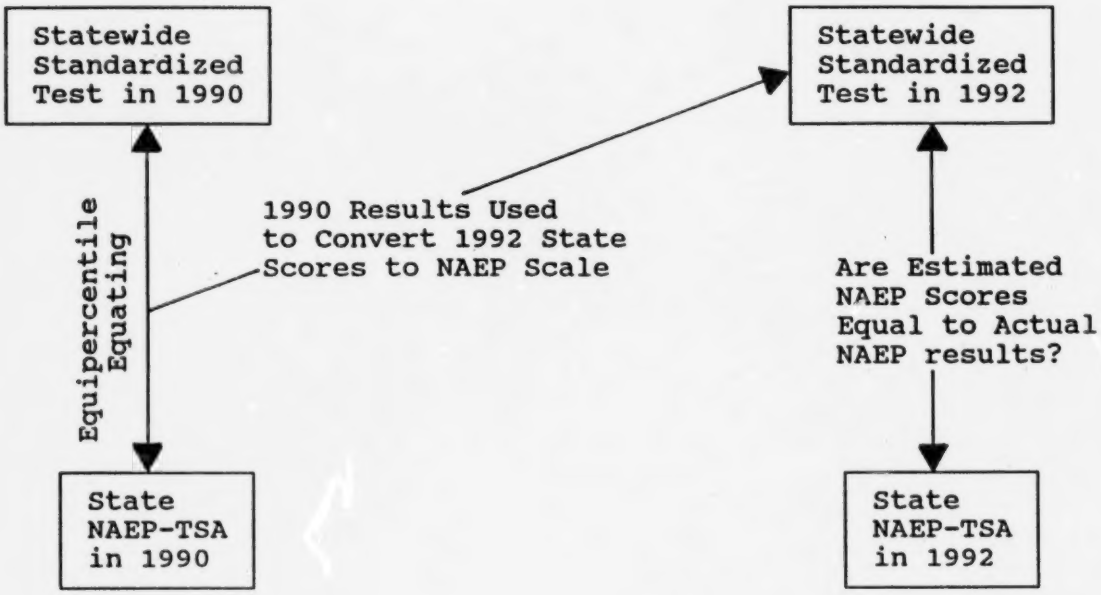
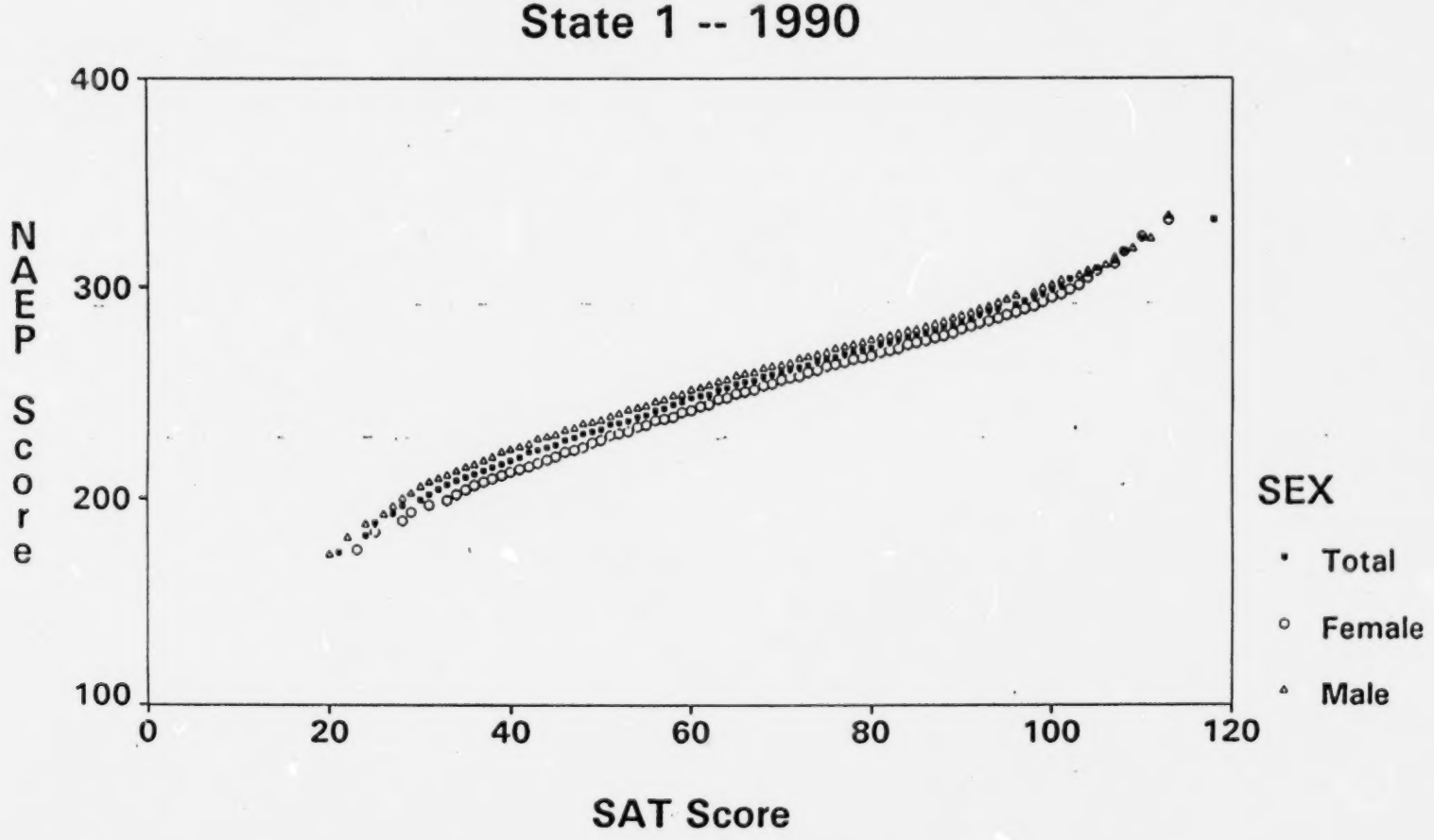

Statewide Standardized Test in 1990
Statewide Standardized Test in 1992
Equipercentile Equating
1990 Results Used to Convert 1992 State Scores to NAEP Scale
Are Estimated NAEP Scores Equal to Actual NAEP results?
State NAEP-TSA in 1990
State NAEP-TSA in 1992

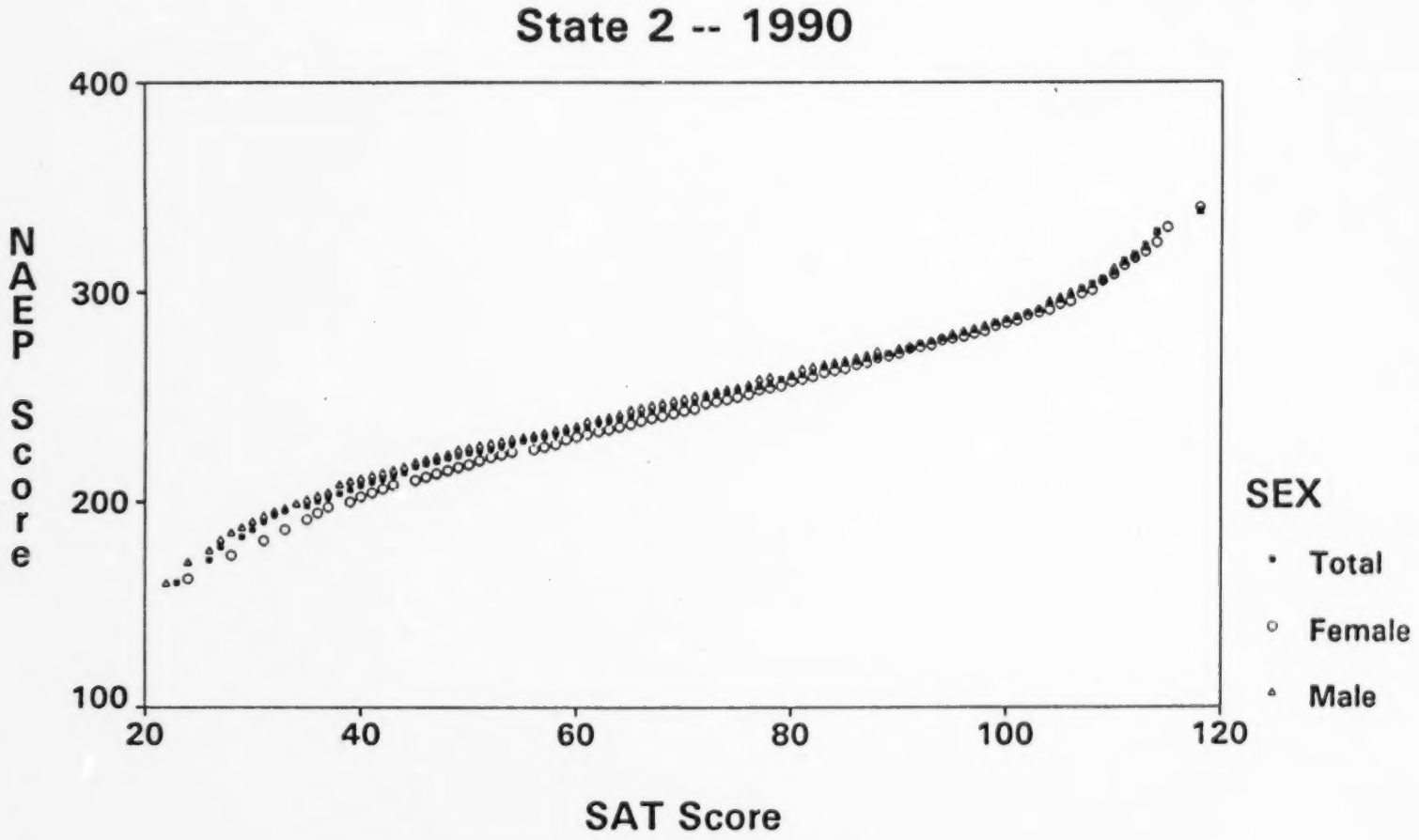
Figure 2. Equating Functions for NAEP Math Scores* and SAT Total Math Scores**



* NAEP scores are average overall mathematics proficiency scores.

**SAT scores are total mathematics scores.

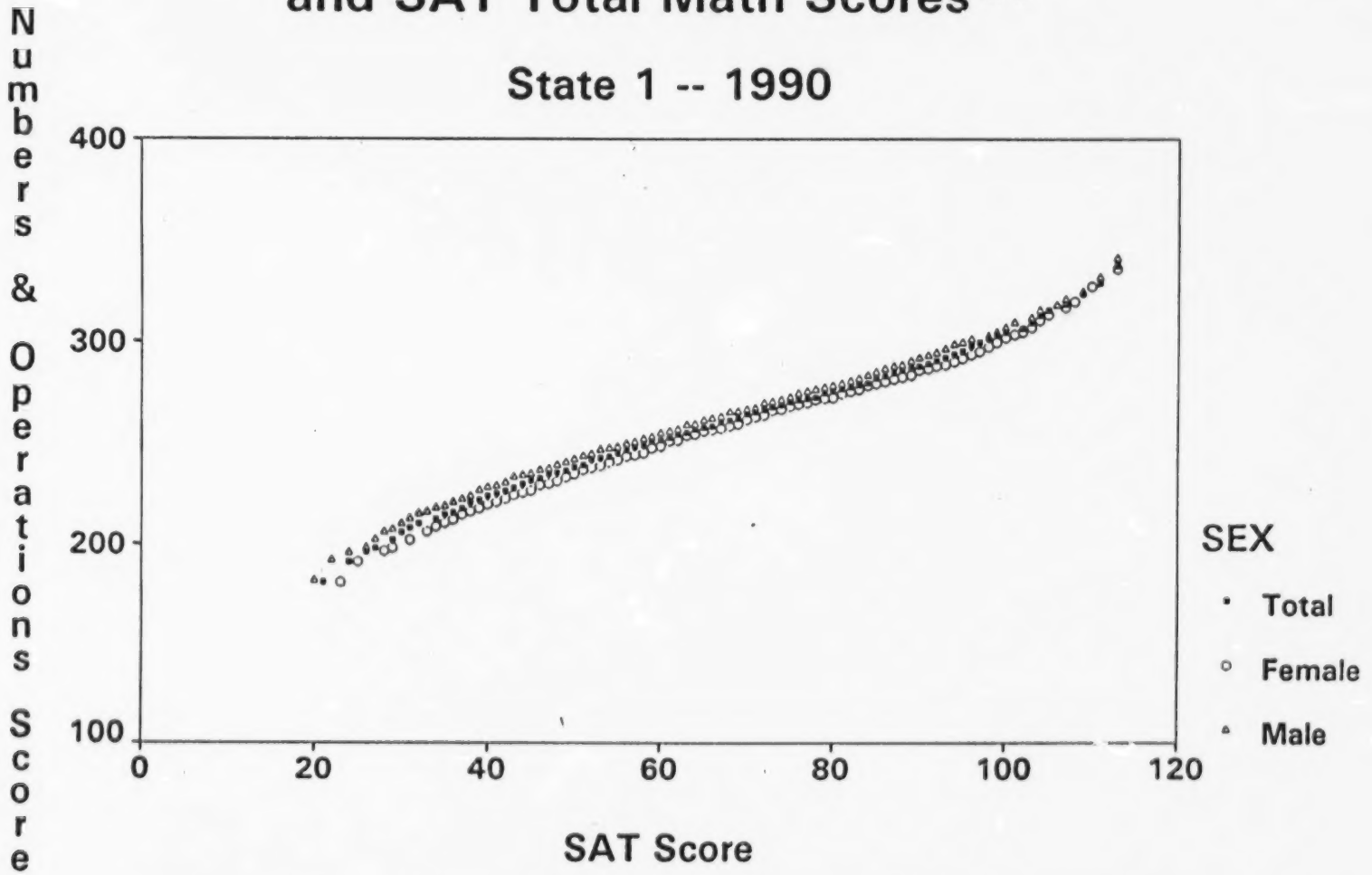
Figure 3. Equating Functions for NAEP Math Scores* and SAT Total Math Scores** -- 1990



* NAEP scores are average overall mathematics proficiency scores.

**SAT scores are total mathematics scores.

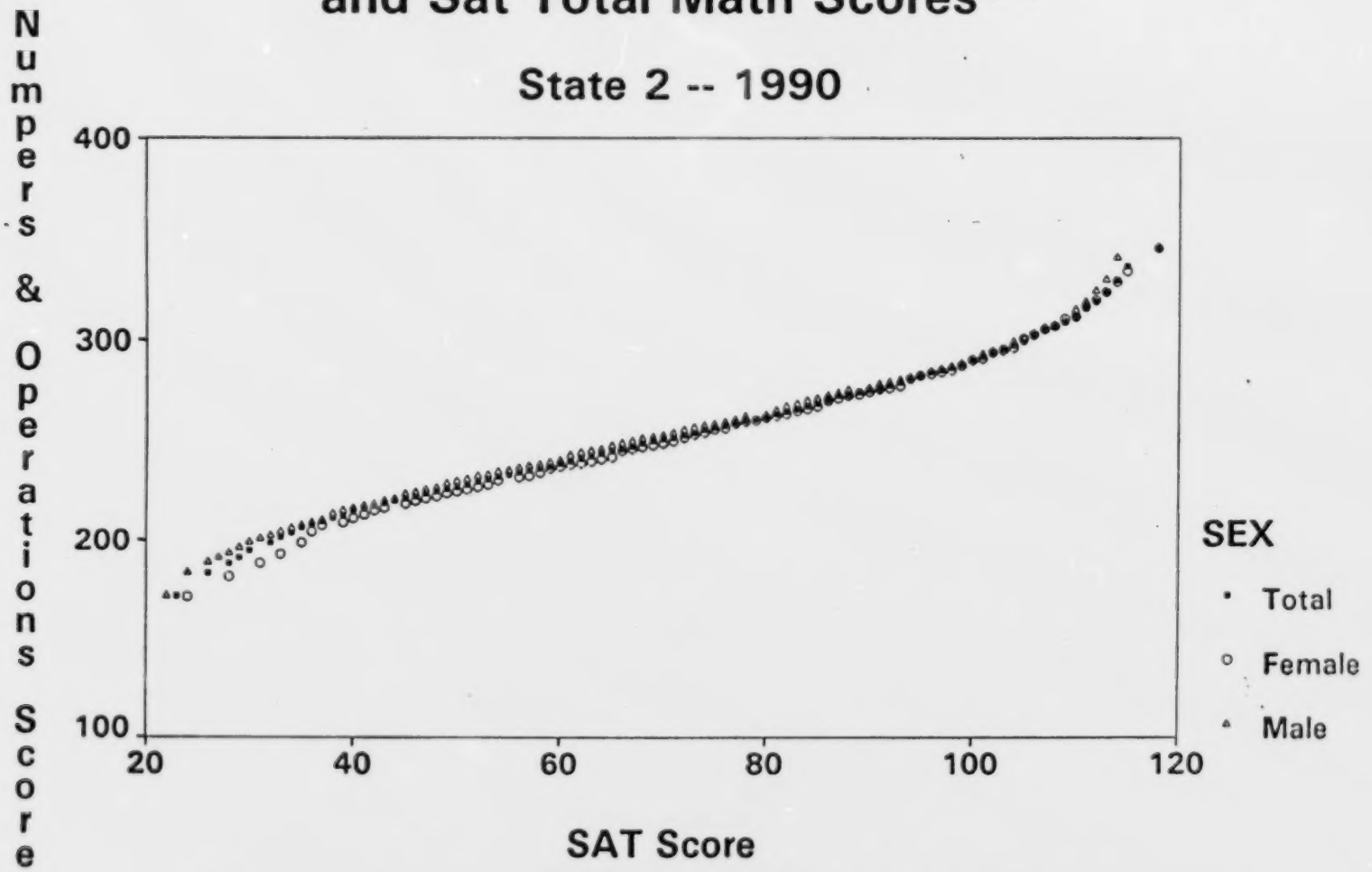
Figure 4. Equating Function for NAEP Math Scores* and SAT Total Math Scores**



* NAEP scores are the 1st plausible value on numbers & operations items.

**SAT scores are total mathematics scores.

Figure 5. Equating Function for NAEP Math Scores* and Sat Total Math Scores**



* NAEP scores are the 1st plausible value on numbers & operations items.

** SAT scores are total mathematics scores.

Table I

Statewide Tests, Forms, and Sample Sizes
for the Four Participating States

Statewide Test Forms				Sample Size	
State	Year	Test	Form	Statewide Test*	NAEP TSA
1	1990	SAT	K	48,991	2,531
	1992	SAT	K	50,413	2,623
2	1990	SAT	E	11,121	2,551
	1992	SAT	L	11,242	2,454
3	1990	ITBS	G	15,309	2,716
	1992	ITBS	G	16,364	2,645
4	1990	CAT	E	73,881	2,843
	1992	CAT	E	80,065	2,769

* Actual numbers of eighth grade students present on the testing date.

Table 2

Scores Corresponding to Selected Percentiles on the SAT for the Total Population and Equivalent NAEP Average Overall Proficiency Scores from Total, Male, and Female Equatings, 1990*

State 1

		Equivalent NAEP Scores			
Total Group Percentile (1)	State Total SAT (2)	State Total (3)	Males (4)	Females (5)	Difference Males Minus Females (6)
95	105	309	311	308	3
90	99	297	299	293	6**
75	84	276	279	273	6**
50	65	253	258	250	8**
25	47	229	233	223	10**
10	34	209	213	202	11**
5	28	196	200	189	11**

* The SAT scores are for Total Mathematics at grade 8 on the Stanford Achievement Test, Form K. The NAEP scores are for the grade 8 Overall Mathematics Proficiency scale based on the average of 5 plausible values.

** Difference greater than twice the standard error.

Table 3

Scores Corresponding to Selected Percentiles on the SAT for the Total Population and Equivalent NAEP Average Overall Proficiency Scores from Total, Male, and Female Equatings, 1990*

State 2

		Equivalent NAEP Scores			
Total Group Percentile (1)	State Total SAT (2)	State Total (3)	Males (4)	Females (5)	Difference Males Minus Females (6)
95	111	315	314	313	1
90	107	302	304	299	2
75	95	279	280	278	2
50	73	251	252	248	4
25	50	224	225	218	7**
10	36	200	203	195	8**
5	30	187	190	186	4

* The SAT scores are for Total Mathematics at grade 8 on the Stanford Achievement Test, Form E. The NAEP scores are for the grade 8 Overall Mathematics Proficiency scale based on the average of 5 plausible values.

** Difference greater than twice the standard error.

Table 4

Scores Corresponding to Selected Percentiles on the SAT for the Total Population and Equivalent NAEP Numbers and Operations Scores from Total, Male, and Female Equatings, 1990*

State 1

		Equivalent NAEP Scores			
Total Group Percentile (1)	State Total SAT (2)	State Total (3)	Males (4)	Females (5)	Difference Males Minus Females (6)
95	105	314	318	313	5
90	99	302	304	300	4
75	84	282	283	279	4**
50	65	259	261	256	5**
25	47	236	237	230	7**
10	34	216	218	209	9**
5	28	204	206	196	10**

* The SAT scores are for Total Mathematics at grade 8 on the Stanford Achievement Test, Form K. The NAEP scores are for the grade 8 Numbers and Operations scale based on the first plausible value.

** Difference greater than twice the standard error.

Table 5

Scores Corresponding to Selected Percentiles on the SAT for the Total Population and Equivalent NAEP Numbers and Operations Scores from Total, Male, and Female Equatings, 1990*

State 2

		Equivalent NAEP Scores			
Total Group Percentile (1)	State Total SAT (2)	State Total (3)	Males (4)	Females (5)	Difference Males Minus Females (6)
95	111	319	319	317	2
90	107	306	306	305	1
75	95	284	283	283	0
50	73	257	257	254	3
25	50	229	230	225	5**
10	36	207	209	204	5
5	30	194	199	188	11**

* The SAT scores are for Total Mathematics at grade 8 on the Stanford Achievement Test, Form E. The NAEP scores are for the grade 8 Numbers and Operations scale based on the first plausible value.

** Difference greater than twice the standard error.

Table 6

Estimated and Actual 1992 NAEP Scores at Selected Percentile Points Based on Equipercentile Equatings of 1990 Stanford Achievement Test Grade 8 Total Mathematics Scores with 1990 NAEP Overall mathematics Proficiency Scores

State 1

Percentile	SAT Score	Estimated Equivalent NAEP Score	Observed NAEP Score	Observed Minus Estimated
95	106	313	311	-2
90	99	297	299	2
75	84	276	276	0
50	64	252	251	-1
25	47	229	227	-2
10	35	210	206	-4*
5	29	199	193	-6*
Mean	65.6	255	251	-4*

* Difference greater than twice the standard error.

Table 7

Estimated and Actual 1992 NAEP Scores at Selected Percentile Points Based on Equipercentile Equatings of 1990 Stanford Achievement Test Grade 8 Total Mathematics Scores with 1990 NAEP Overall mathematics Proficiency Scores

State 2

Percentile	SAT Score*	Estimated Equivalent NAEP Score	Observed NAEP Score	Observed Minus Estimated
95	110	307	317	10**
90	106	295	305	10**
75	94	275	283	8**
50	76	252	257	5**
25	55	239	231	2
10	39	210	208	-2
5	31	197	194	-3
Mean	79.8	256	257	1

* SAT Form L (used in 1992) equated to SAT Form E (used in 1990).

** Difference greater than twice the standard error.

Table 8

Estimated and Actual 1992 NAEP Scores at Selected Percentile Points Based on Equipercentile Equatings of 1990 Iowa Test of Basic Skills Grade 8 Total Mathematics Scores with 1990 NAEP Overall Mathematics Proficiency Scores

State 3

Percentile	ITBS Score	Estimated Equivalent NAEP Score	Observed NAEP Score	Observed Minus Estimated
95	196	318	323	5*
90	189	307	313	4*
75	179	292	296	4*
50	166	271	275	2
25	154	251	254	2
10	142	231	235	2
5	136	218	223	1
Mean	166	271	274	3*

* Difference greater than twice the standard error.

Table 9

Estimated and Actual 1992 NAEP Scores at Selected Percentile Points Based on Equipercentile Equatings of 1990 California Achievement Test Grade 8 Total Mathematics Scores with 1990 NAEP Overall Mathematics Proficiency Scores

State 4

Percentile	CAT Score	Estimated Equivalent NAEP Score	Observed NAEP Score	Observed Minus Estimated
95	838	320	315	-5*
90	827	308	303	-5*
75	806	286	282	-4*
50	783	258	258	0
25	760	233	234	1
10	738	214	212	-2
5	722	202	199	-3*
Mean	782	256	258	2

* Difference greater than twice the standard error.